Podatkovno rudarjenje iz časovne vrste ozon - napoved brez napovednikov

Boris Bizjak
Univerza v Mariboru, FERI
Smetanova 17, 2000 Maribor
boris.bizjak@um.si

Data mining ozone - forecasting without predictors

Data mining is the computational process of discovering patterns involving methods at intersections of artificial intelligence, machine learning, statistic and database. In this paper we discus air pollution methods forecasting: Data Mining at ARMA (Autoregressive and Moving Average Models) and Data Mining at ART (Autoregressive Tree Models).

Many areas experience elevated concentrations of ground-level ozone pollution during the summertime "smog season". Local environmental or health agencies often need to make daily air pollution forecasts for public advisories and for input into decisions regarding abatement measures and air quality management. Such forecasts are usually based on statistical relationships between weather conditions and ambient air pollution concentrations. We use only ozone time series data in forecast algorithm. This work showed real time daily maximum ozone forecasting performance at two monitor sites Nova Gorica and Koper, Slovenia. Six month results at Koper (2014 march – 2014 august) shows good forecasting: RMSE 12% at first day forecast, RMSE 16% at second day and RMSE 17% at third day forecast. At measure points Koper we get only 93% of all time series data. Algorithm parameter "missing value substitution" is previous.

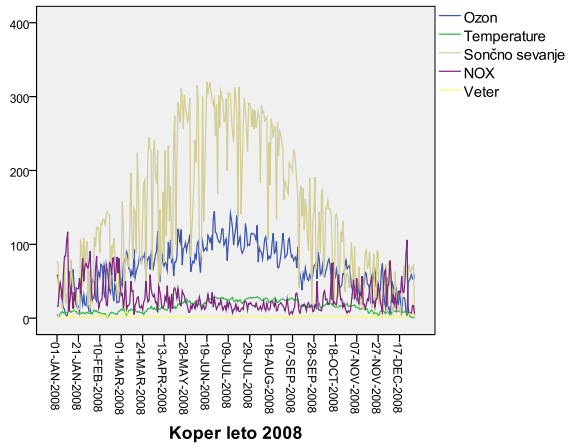
1 Uvod

V članku predstavljamo samodejno napoved O_3, kot primer uporabe podatkovnega rudarjenja. Povišane koncentracije ozona v nizkih plasteh ozračja so posledica onesnaženja zraka z NO_x in HC, kar je posledica prometa in industrije. Ozon nastaja tudi v naravnem okolju[1]. Visoke koncentracije se pojavljajo poleti, ko so visoke vrednosti sončnega sevanja. Visoke vrednosti O_3 imajo negativne vplive na človeško zdravje, posebej na dihalne poti, kašelj in poslabšanje pljučnih funkcij. V obdobju visokega ozona se poveča umrljivost ljudi. Ozon je močan oksidant in kot takšen vpliva na ljudi in rastline. Napovedovanje ozona ima smisel od marca do septembra.

Modeli igrajo pomembno vlogo v analizi in pojasnjevanju inženirskih, sociooloških, ekonomskih, in medicinskih fenomenov. Mnogokrat se srečujemo s fenomeni, ki jih ne moremo opisati z enostavno razumljivimi strojnimi modeli (deterministični modeli). Koncentracija ozona (y) je odvisna od temperature (x_1), sončnega sevanja (x_2), koncentracije NO_x (x_3) in vetra (x_4)(enačba 1).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \quad (1)$$

Enostavno napoved koncentracije lahko izvedemo z uporabo linearne regresije. Ta princip smo opisali v članku na enih starejših konferenc ERK. Kasneje smo linearno regresijsko metodo napovedi adaptirali z nelinearnim vplivom vetra. Točnost napovedi je bila solidna, slabost je bila odvisnost od točnosti vremenske napovedi smeri in jakosti vetra.



Slika 1: Izmerjeni podatki za zrak Koper leto 2008

Slika 1 prikazuje nelinearne in sezonske odvisnosti med posameznimi vplivnimi faktorji za napoved O_3. Smiselno je linearno regresijsko metodo napovedi zamenjati z nelinearno metodo napovedi. Taki metodi ali metodologiji sta ART [5] (Autoregressive Tree Models) in ARMA[4][6] (Autoregressive and Moving Average Models). ART in ARMA metodologijo lahko pojasnimo na več načinov, tudi s teorijo digitalnih filtrov [7], torej sta metodologiji spektralni metodi ali z drugi besedami, v osnovi za dobro napoved ne potrebujeta dodatnih amplitudnih napovednikov k osnovni časovni vrsti [3][4]. Uporabili smo strukturo podatkovno rudarjenje (ang. Data Mining) [3], ki jo sestavljajo baza podatkov, strojno učenje in statistika. Za več dnevno napoved izvajamo strojno učenje enkrat dnevno, za urno napoved vsako uro. Osnovno strukturo algoritma napovedi smo preverjali in

določili s simulacijskim programom podatkovnega rudarjenja.

2 Uporabljene metode

Autoregressive model AR(p). Večina časovnih vrst je sestavljena iz elementov, ki so medsebojno odvisni. Da bi laže razumeli avtoregresijski proces: prvi in najpomembnejši korak je transformacija osnovne časovne vrste iz enega stolpca, v tabelo z več stolpci – v več novih spremenljivk (tabela 1). Tako element "danes" opišemo s serijo specifičnih elementov iz preteklosti. To imenujemo avtoregresijski proces in ga opišemo z enačbo:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \qquad (2)$$

φ_i……..parameter modela,
c……konstanta,
ε_t ……..naključna napaka (ang. random error),
p……red modela (ang. model order).

Moving average model MA(q). Neodvisno od avtoregresijskega procesa, vsak element je lahko odvisen tudi od napake iz preteklosti:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \qquad (3)$$

θ_i……parameter modela,
$\varepsilon_t, \varepsilon_{t-i}$……napaka,
q……red modela (ang. model order),
μ…….srednja vrednost časovne vrste.

Večina časovnih vrst ima sezonski vzorec. Za dobro napoved je potrebno, da algoritem razume periodičnost podatkov – sezonski vzorec. Za periodičnost algoritem generira dodatne vhodne spremenljivke. Za primer, če so podatki mesečni, torej periodični z 12, algoritem doda dodatne spremenljivke Max (-12), Max (-24) itd. Za avtomatsko določitev periodičnosti lahko uporabimo Fourierov transform (FFT).

Mesec	Max
Jan 08	464
Feb 08	675
Mar 08	703
Apr 08	887
Maj 08	1139
…	…

ID	Max (-2)	Max(-1)	Max (0)
1	464	675	703
2	675	703	887
3	703	887	1139
…	…	…	…

Tabela 1 Transformacija osnovne časovne vrste v tabelo za izvedbo avtoregresijskega procesa

Avtoregresijsko drevo je kombinacija avtoregresijskega procesa in odločitvenega drevesa. Vsi stolpci Max -1, Max -2, ….Max(-12), Max(-24) so vhodne spremenljivke (ang. regresor), stolpec Max(0) je napovedan izhod. Naloga inteligence ART drevesa je določiti, katere vhodne spremenljivke so smiselne za določitev napovedi. Vpliv odločitvenega drevesa je tudi, da linearne regresijske odvisnosti razreže v točkah nelinearnosti. Pri časovnih vrstah pride to v poštev tudi ob vsaki časovni spremembi karakterističnega vzorca v podatkih. Algoritem podatkovnega rudarjenja spremlja spremembe karakterističnega vzorca vhodne časovne vrste, kar se odraža v spremembi ART drevesa. ART algoritem je optimiziran za napoved naslednjega koraka v časovni vrsti (Max+1). Za naslednji korak napovedi (Max+2) tvorimo novo ART drevo, itd. Tako postane ART metoda za dolgoročne napovedi počasna in nepraktična. V teoriji »se sliši redko«, a v praksi napovedane vrednosti hitro postanejo nestabilne, pomeni: napovedane vrednosti eksponentno narastejo ali upadejo, ali celo divje zanihajo. Zaradi tega smo uporabili dve metodi za napovedovanje: ART in ARMA. ART je stabilna za kratkoročne napovedi in postane nestabilna za dolgoročne napovedi, ARMA je bolj stabilna za dolgoročne napovedi.

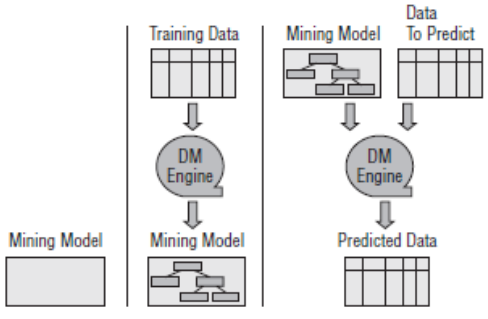
Enostavni nazorni primer uporabe opisanih metod je model napovedi za tri dni, za 27.8.2014, 28.8.2014 in 29.8.2014 (Slika 4). Za učenje modela napovedi smo uporabili časovno vrsto 182 vrednosti Max O3, ki so prikazane na sliki 5. Časovna vrsta 182 Max vrednosti je izračunana iz 4300 urnih meritev, z SQL agregatno funkcijo MAX.

Enačba ART drevesa algoritma napovedi:
Max Vrednost = 25,9023520330278
+ 0,764938744642134 * Max Vrednost(-1)

ARIMA enačba algoritma napovedi:
({1},0,{1,0.735598539528764,0.444109955866561,0.329272181713964,0.243912110006262,0.213583085866147,9.79190035657538E-02,0.106895954009619})
X ({1},1,{1,-0.769584282482033})(16)
Intercept:0.934730575470153

ARIMA enačba je prikazana v multiplikativni obliki. Končna ARIMA enačba se izračuna z multiplikacijo izrazov.

3 Proces podatkovno rudarjenja



Slika 3: Blokovna ponazoritev procesa podatkovnega rudarjenja

Podatkovno rudarjenje je interdisciplinarno pod-področje računalniške znanosti, ki išče vzorce v velikih količinah podatkov z uporabo metod: umetna inteligenca, strojno učenje, statistika in podatkovna skladišča. Glavni cilj podatkovnega rudarjenja je izluščiti informacije iz podatkov v razumljivo obliko, v primerno obliko za kasnejšo uporabo. Osnovna ideja podatkovnega rudarjenja je, da algoritem avtomatsko izlušči karakteristični vzorec iz vzorčnih podatkov, vzorec nato uporabi za izračun npr. napovedi.

Arhitekturo podatkovnega rudarjenja smo sestavili iz podatkovnega skladišča (MS SQL Server) in pripadajočega strežnika za analize (MS Analysis Server). Vizualizacijo podatkov smo izvedli z WEB strežnikom (MS IIS7, NET framework 4). Server za analize komunicira s podatkovnim strežnikom preko XMLA protokola (ang. XML for Analysis). XMLA je industrijski standard za prenos podatkov v sistemih za analizo, je XML protokol prirejen za komuniciranje s strežniki za analize, format je neodvisen od izvora ali prejemnika podatkov.

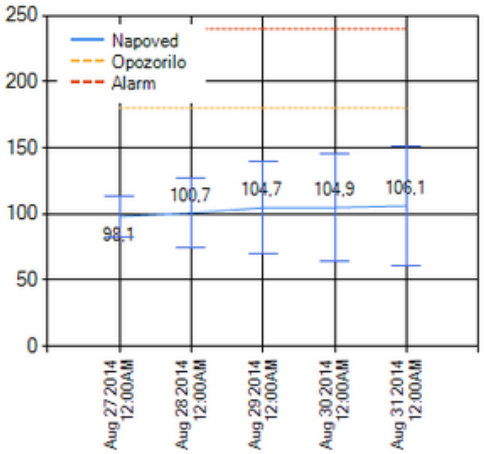
4 Rezultati - kratkoročna in dolgoročna napoved ozona

Izvedli smo napovedi za ozon za Koper in Novo Gorico, za dva merilna mesta na različnih geografskih lokacijah. Empirična modela napovedi za omenjani merilni točki se razlikujeta, a smo bili podobno uspešni za obe merilni točki, o čemer pričajo cenilci kakovosti modelov napovedi: MAE - absolutni pogrešek napovedi, MAPE - relativni pogrešek napovedi, RMSE - standardna deviacija pogreškov napovedi.

Dan napovedi	MAE	MAPE	RMSE	Število napovedi
5+	22,9	19,4 %	27,5	143
4+	19,7	18,6 %	26,5	142
3+	19	17,4 %	25,2	141
2+	17,5	15,8 %	22,5	139
1+	14,1	12,7 %	18,2	105

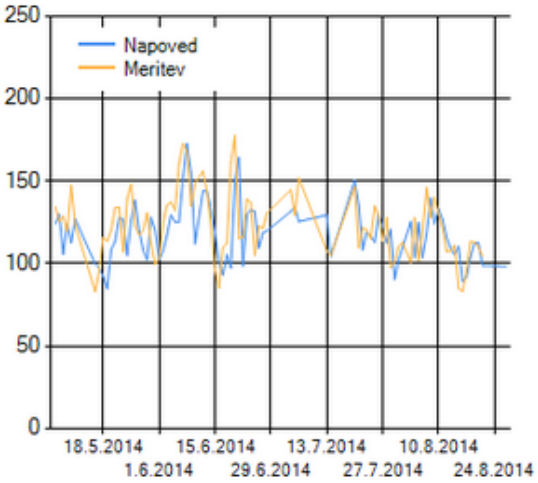
Tabela 2: Kakovost modela napovedi za 5 dni O_3 - Koper od 1. marec 2014 do 21.7.2014

Zapišimo nekaj zanimivih izsledkov iz letnega dela. Prve 5 urne napovedi smo izvedli, kot se reče »na vroče«, brez simulacij, to je relativno enostavno, saj počakate nekaj ur in vidite rezultate. Za napredne 5 urne optimizacije napovedi ali 5 dnevne napovedi, smo izdelali simulacijski program. Takšen simulacijski program je nekoliko drugačen, kot jih poznamo iz IBM SPSS ali MatLAB. Osnovni algoritem simulacije poteka tako, da se pomikamo po bazi podatkov, za en korak. Po vsakem pomiku po bazi izvedemo učenje, nato izvedemo napoved, rezultate napovedi pa shranimo v nazaj v bazo podatkov, in tako naprej za vse historične podatke, za katere želimo izvesti preverbo ideje modela napovedi. Z simulacijo ugotavljamo, npr. kako ustrezno razporediti uteži med ARMA, ART in ARMA sezonski. Stopnja ARMA in ART se določi avtomatsko s strojnim učenjem.



Slika 4: Napoved z deviacijo za 5 dni - dnevni urni ekstremi O3 - Koper avgust 2014

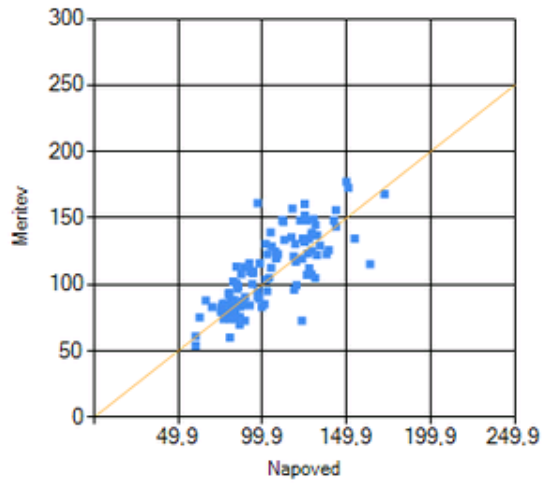
Strojno učenje in napovedi in se izvajajo dinamično, za primer urne kratkoročne napovedi vsako uro, za primer napovedi dnevnih ekstremov 1 krat dnevno. Takrat preračunavamo tudi MAE, MAPE in RMSE (tabela 2). Rezultati prikazani v članku, so dostopni tudi kot dinamična WEB aplikacija, vendar bodo tam rezultati ob vašem obisku verjetno drugačni, ker je to aplikacija realnega časa: http://193.95.233.105/econova5/5dni_O3_final_KP_01.aspx



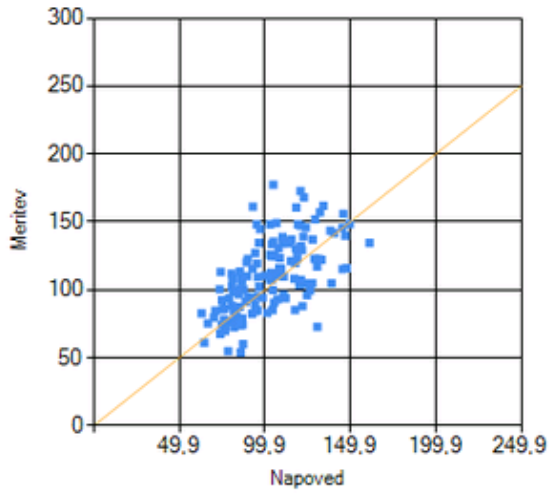
Slika 5: Potek napovedi in meritev dnevnih urnih ekstremov O_3 - napoved za naslednji dan Koper 2014

V kolikor so rezultati simulacij dobri, enostavno izvršimo preklop na realni čas in dobimo WEB aplikacijo, ki deluje v realnem času. Takšen kompleten cikel simulacije za urne napovedi, za dva meseca, traja nekje med 2 in 3 ure. Rezultate, ujemanje napovedi in meritev, za 5 urne napovedi si lahko ogledate na: http://193.95.233.105/econova5/default_nova.aspx.

Simulacijski program za dnevne urne ekstreme, se za 6 mesecev izvaja 1 uro. Krajši čas simulacije smo dosegli tako, da smo procesirali samo časovno vrsto dnevnih urnih ekstremov, in ne vseh merilnih podatkov. Vsekakor je za nekaj dni naprej teže napovedati, kot pa nekaj ur – potrebnega je več dela, več simulacij. Modeli za urne in dnevne napovedi so strukturno različni.



Slika 6: Ujemanje med meritvijo in napovedjo urni max 1 dan naprej - Koper



Slika 7: Ujemanje med meritvijo in napovedjo urni max 2 dni naprej – Koper

Večji problem (saj želimo čim boljšo napoved) so bili manjkajoči merilni podatki, ki jih zagotavlja ARSO. Manjkajoče merilne podatki smo nadomeščali s »prejšnjim« merilnim podatkom (ang. previous) ali z srednjo vrednostjo meritev (ang. mean), a to nadomeščanje deluje dobro le do neke mere. Algoritmi strojnega učenja se sami odzovejo na »luknje« v merilnih podatkih. Za časovne vrste z manjkajoči podatki optimalni modeli napovedi samodejno postanejo enostavni (npr. avtoregresijsko drevo je nižje stopnje), kar pomeni slabše napovedi.

V sistemih je težko izmeriti podatke za napovednike (ang. predictors) ali se dokopati do ustreznih baz podatkov, na razpolago so samo historične časovne vrste osnovnega pojava – v našem primeru koncentracije ozona. Kljub temu smo poizkusili modele napovedi izboljšati z delnim upoštevanjem meteorologije. Rezultati so nekoliko boljši, vendar ne drastično. Izračuni so pokazali, da ima smisel upoštevati le dva napovednika: smer vetra in vlažnost zraka. Takšen hibridni model je žal tudi bolj občutljiv na izpade merilnih podatkov, saj potrebuje za normalno delovanje 3 kompletne časovne vrste in ne več samo ene: časovne vrste O_3.

Zaključek

Izkustveni modeli napovedovanja se gradijo »po naročilu«, ni fiksnega recepta. Vprašanje pri napovedovanju je, kako določiti pravo strukturo modela in nato parametre. To je čar napovedovanja, potrebno je eksperimentirati in rezultati so pogojeni tudi z intuicijo raziskovalca. Pri svojem delu smo upoštevali izsledke EPA (ang. United States Office of Air Quality) [1] in dodali smo najnovejšo informacijsko tehnologijo podatkovnega rudarjenja. Rezultate smo obdelali po metodologijah, ki so običajne za takšne raziskovane vsebine [2]. Primerljivost z ostalo literaturo je težavna, saj se za vsako merilno točko izmeri svoja časovna vrsta in izračuna svoja struktura modela napovedi in je t.i. »cross prediction« smiselno težko izvedljivo. Potrdili smo, da sta ART in ARMA metodologiji napovedovanja, ki z najmanj vhodnimi podatki ponudita največ.

Literatura:

[1] EPA, Guideline for Developing an Ozone Forecasting Program, United States Office of Air Quality EPA-454/R-99-009 Environmental Protection Planning and Standards July 1999

[2] Luis A. Dı´az-Robles, Juan C. Ortega, Joshua S. Fu b, Gregory D. Reed b, Judith C. Chowc, John G. Watson c, Juan A. Moncada-Herrera. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile, Atmospheric Environment, 42 (2008), ELSEVIER

[3] Jamie MacLennan, Zhao Hui Tang, Bogdan Crivat. Data mining with Microsoft SQL Server 2008, Wiley

[4] GEORGE E. P. BOX GWILYM M. JENKINS GREGORY C. REINSEL. Time Series Analysis Forecasting and Control, Wiley

[5] C.Meek, D.M. Chickering, D.Heckerman, Autoregresive Tree Models for Time-Series Analysis, Microsoft Research

[6] Bo Thiesson, David Maxwell Chickering, David Heckerman, Christopher Meek, ARMA Time-Series Modeling with Graphical Models, Microsoft Research

[7] Leland B. Jackson, Digital filters and Signal Processing, Kluwer academic publishers, 2002, Fifth Printing